Patterns of Codon Usage Bias in WRKY Genes of *Brassica rapa* and *Arabidopsis thaliana*

Snigdha Srivastava[1], Sheetal Chanyal[1], Ashutosh Dubey[2], A. K. Tewari[3] & Gohar Taj[1]

[1] Department of Molecular Biology and Genetic Engineering, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

[2] Department of Biochemistry, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

[3] Department of Plant Pathology, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

Correspondance: Gohar Taj, Department of Molecular Biology and Genetic Engineering, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India. E-mail: gohartajkhan@rediffmail.com

Abstract

Codon usage bias (CUB) is defined as the selective and nonrandom use of synonymous codons by the organism for encoding the amino acids. One of the important plant transcription factor family is the 'WRKY' whose role has been investigated in the regulation of abiotic and biotic stress responses in plants. In this paper, the codon usage pattern of the WRKY transcription factor of the two important plant species *Arabidopsis thaliana* and *Brassica rapa* has been investigated. Various codon usage indices like ENc, CAI, correspondence analysis, RSCU analysis, neutrality plot and hierarchial clustering has been done. The GC codon status was high in Arabidopsis. The RSCU analysis of codons revealed that codons coding for arginine was maximum in both the plant species. Our results propose that natural selection was the main dominating factor guiding the evolution of different WRKY genes in both *Arabidopsis thaliana* and *Brassica rapa*.

Keywords: codon usage bias, correspondence analysis, gene expression, hierarchical clustering

1. Introduction

The genetic code is the set of codons used by the living cells to convert the information encoded by the DNA into the proteins. When all the codons coding for the same amino acid are used by the same frequency the term synonymous codon is used. However, an unequal usage frequency for different synonymous codons has been observed. This unequal usage frequency of synonymous codons is termed as codon usage bias (CUB). The reason behind the differences in codon usage is the variation occurring in the choice between codons ending with C or G/A or T (Campbell & Gouri, 1990). CUB throws light on the origin of genes, species and the mutational forces acting on them (Wu et al., 2017) along with it also guides in predicting the functions of related genes, structure of protein and expression of protein (Zhao et al., 2016) (Song et al., 2015). The pattern of codon usage depends on the mutation pressure, natural selection and on the sequence of amino acids (Mandlik et al., 2014). The two widely accepted theories of CUB are neutral theory and selection-mutation-drift theory (Bulmer, 1991). The neutral theory states that mutations at degenerate coding positions are neutral and the synonymous codon choice is random. The other model says that codon bias is aided by a balance between mutation pressure and genetic drift (Yang et al., 2015). There are some other possible factors affecting the CUB within the species. These include gene length (Duret & Mouchiroud, 1999), gene expression level (Hambuch & Parsch, 2005), GC content (Hu et al., 2007), environmental stress (Goodarzi et al., 2008), RNA stability (Akashi, 1997), population size, recombination rate and codon position (Behura & Severson, 2012). Over the past few years with the advent of DNA sequencing technologies, a lot of plant genes have been sequenced and deposited in the databases. This has resulted in the increased study of CUB patterns of different plants and its genes (Gustafsson et al., 2004). Variation in CUB is a unique characteristic of the genome and between specific genes of the species (Supek & Vlahovicek, 2005). Differences in codon usage have also been found in genes within the species. In studies carried out on monocot and dicot species when the third codon position is modified with G or C an increased expression of modified genes was observed. These results give an idea about the relation between the translation

efficiency and codon bias in monocots and dicots (Kawabe & Miyashita, 2003). Codon usage tells about the specific pattern of gene expression and it has also been noted that the genes expressed under the same physiological state prefer the use of same codons (Chiapello et al., 1998). Plants are subjected to a wide variety of stresses simultaneously as they are sessile life forms. During the course of evolution physiological and biochemical adaptations developed in plants provides an advantage to combat a single stress but not multiple. Perception of the external stimulus is sensed by the receptors on the cell membrane that triggers the chain of molecular/biochemical reactions. WRKY are a large family of regulatory proteins which regulate diverse response against biotic and abiotic stresses through a complicated network of genes (Smith, 2000). WRKY transcription factors regulate the gene expression through activation and repression of W-box and W-box like sequences. A tight regulation is involved in the binding of WRKY proteins to the regulatory elements, as a result, they have become a target for crop improvement (Phukan et al., 2016). A possible cause for the expansion of the WRKY transcription factor family during the course of evolution is the exposure to a number of biotic and abiotic stresses (Eulgem et al., 2000). The study herein is focused to analyze the codon bias and base composition in WRKY transcription factor genes of *Brassica rapa* and *Arabidopsis thaliana* using different codon usage indices. This study will improve our understanding of WRKY gene usage pattern as well as there pattern of evolution and function in two plant species.

2. Materials and Methods

2.1 Gene Sequence

Complete coding DNA sequences of WRKY genes of *Brassica rapa* and *Arabidopsis thaliana* were taken from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/). The presence of start codon and a stop codon at the end of the sequence was ensured. Sequences were also checked for the absence of stop codon in the coding frame. The length of the sequences only greater than 300 base pairs were considered in the study.

2.2 Analysis of the Base Composition

The program CAIcal (genomes.urv.es/CAIcal/) was used to calculate the GC content at the first, second and third nucleotide position (GC1, GC2 and GC3) respectively. GC12 is the average value of the GC content at the first and second nucleotide position. The values of GC1, GC2, GC3 and GC12 of *Arabidopsis thaliana* and *Brassica rapa* are mentioned in Tables 1 and 2 respectively.

2.3 RSCU Analysis

The relative synonymous codon usage (RSCU) value for a codon 'i' is defined as the value representing the ratio between the observed usage frequency of one codon in a gene and the expected usage frequency in the synonymous codon family. It is represented as,

$$RSCU_i = obs_i/exp_i \tag{1}$$

where, obs_i is the observed number of occurrences of codon i and exp_i is the expected number of occurrences of the same codon (based on the number of times the relevant amino acid is present in the gene and the number of synonymous alternatives to i). The pattern followed for synonymous codon usage was assessed by RSCU analysis discrediting the influence of amino acid composition. This index reflects the relative usage preference for a specific composition of codons encoded in the same amino acid (Wang et al., 2016). The RSCU of the WRKY sequences of *Arabidopsis thaliana* and *Brassica rapa* were calculated using MEGA 7 software by excluding the stop codons and the codons which code amino acid by a single code. Given that all codons for the particular amino acid are used equally. The RSCU value > 1 depicts positive CUB, value < 1 depicts negative CUB and 1 represents no CUB (Sharp & Li, 1986). The RSCU value > 1.6 shows strongly preferred codons.

2.4 ENc Analysis

The CUB present in a gene were calculated using ENc (Effective number of codon usage). The value of ENc ranges from 20 (representing extreme bias where one codon for each codon family is used) to 61 (representing random usage of all synonymous codons). The program used for the ENc calculation was genomes.urv.es/CAIcal/. The ribosomal protein gene of *Arabidopsis thaliana* and *Brassica rapa* was used as a reference set.

2.5 Codon Adaptation Index Analysis

CAI (Codon Adaptation Index) is another widely used method of CUB. It measures the deviation present in the given protein-coding gene sequence with respect to a reference set of genes. The program genomes.urv.es/CAIcal/ was used for CAI calculation (Puigbò et al., 2008). The range of CAI value is from 0 to 1. The synonymous codon usage pattern of ribosomal genes was used as a reference set. In order to ascertain the

relative influence of mutation and selection, the correlation analysis between CAI and ENc values was carried out. If the selection is given more preference over mutation, the value of correlation (r) between the two quantities should be very high($r \to -1$). Contrastingly, if mutation force is more important, r approaches 0 (no correlation) (Chen et al., 2014).

2.6 ENc and GC3 Plot

ENc versus GC3 plot were used to analyze whether the codon usage in a set of genes is affected by mutation, selection and by other factors. When genes are distributed along the standard curve it is affected by mutation and if fall below the selection pressure operates on codon usage. The formula used in the study for calculation of expected ENc value and actual ENc value was described by Wright in 1990. The formula for actual ENc value is $2 + (9/F2) + (1/F3) + (5/F4) + (3/F6)$, where, F2, F3, F4, F6 is the average homozygosity of the codons with 2, 3, 4, and 6 synonymous codons respectively. Expected ENc value was calculated by the formula: $2 + GC3 + 29/(GC3)^2 + (1 - GC3)^2$.

2.7 Multivariate Statistical Analysis for RSCU Values of WRKY Genes

One of the most commonly used statistical approach to analyze synonymous CUB is correspondence analysis (Greenacre, 1984) which was implemented in the NPSS 12 software. Correspondence analysis (COA) was applied on RSCU values to investigate the trend used by WRKY genes in the two plant species, *Arabidopsis thaliana* and *Brassica rapa*. In this analysis for minimization of the effect of amino acid composition on codon usage, each WRKY gene of the two plant species was plotted as a 59-dimensional vector space. Each dimension corresponds to the RSCU value of one sense codon. The three termination codons along with codons of methionine and tryptophan were excluded. The trend of variation among the genes can be observed from the measures of relative inertia. WRKY genes were finally ordered according to the position along the axis of major inertia.

2.8 Hierarchial Clustering of WRKY Genes

We used the 59 RSCU values of each WRKY gene from the two plant species for there systematic classification. Clustering is a technique that creates clusters of data points closer to each other, and also far apart from data points in other clusters. The heat map and phylogenetic tree of different WRKY genes of the two species were constructed using NCSS 12 software clustered heat map method. The relationship of RSCU with the phylogeny was further analyzed. Each coding WRKY gene sequence was considered as a separate class, and then according to the distance between these sequences, two sequences that have the minimum distance are merged into a single class.

2.9 Neutrality Plot

The two factors affecting the CUB are mutational pressure and natural selection. The extent to which mutational pressure affects CUB as compared to selection pressure were determined by neutrality plot analysis (Sueoka, 1988). The occurrence of synonymous codon mutation is at the third position of the codon, but sometimes mutations may also occur in the first and second positions resulting into non-synonymous codons. A graph plotted using GC_{3s} on the X-axis and GC_{12} on the Y axis and further calculation of regression line determined mutation-selection equilibrium coefficient. The regression curve effectively measures the degree of neutrality, regression line that falls near the diagonal (slope = 1) shows weak selection pressure on the CUB, whereas deviation of the regression curve from the normal indicate the large influence of natural selection on CUB (Kumar et al., 2016).

3. Results

3.1 Analysis of the Base Composition

The AT and GC content of a gene have an important role to play in gene organization. The GC rich plant genes help the plant to respond to environmental stress (Tatarinova et al., 2010). Important features of the GC base pair are its higher mutability because cytosine gets frequently methylated. (Ossowski et al., 2010; Coulondre et al., 1978) and more cost of synthesis as compared to AT pair. In the present study, the AT content was higher than GC content in each of the WRKY gene CDS examined in both the plant species. Although optimal codons were mostly found to end with G/C. Song et al. (2015) analyzed the WRKY genes of *Glycine max* and reported higher AT content as compared to GC content. AT content in *Arabidopsis thaliana* ranges from 50 % to 65% and of *Brassica rapa* ranges from 50% to 58%. The results showed that the G+C content at the three codon positions was noticeably different GC1 is higher than GC2, and GC3 was lowest in all the three positions in both the plants WRKY genes as shown in (Tables 1 and 2). The low GC3s content in *Arabidopsis* genes is consistent with the previous reports (Tatarinova et al., 2010). After comparing the G and C ending status of codons of WRKY

genes of *Brassica rapa* and *Arabidopsis thaliana* it was seen that the status of G and C ending codons was high in *Arabidopsis thaliana* genes as compared to *Brassica rapa* genes. A very high G and C ending status were present in some of the WRKY genes of two species. Out of the 70 WRKY genes analyzed from both the species, 15 WRKY genes of *Brassica rapa* and 30 WRKY genes of *Arabidopsis thaliana* showed codons with more frequent G and C ends. There is the importance of codons ending with G and C ends in determining protein functions like it was observed that codon usage of genes encoding for regulatory proteins naming transposases, kinases, transcription factors, and phosphatases are biased towards G and C ending codons (Fennoy & Bailey-Serres, 1993). The status of CG and TA doublet codons were higher in *Brassica rapa* WRKY genes namely 1, 2, 8, 9, 10, 13, 15, 18, 21, 25, 32, 33, 44, 46, 47, 48, 53, 54, 56, 58, 59, 62, 64, 69, 70. The doublet TA is a least preferred combination at most of the second and third codon position in most of the eukaryotes. In our study, it was seen that in some of the WRKY genes of *Arabidopsis thaliana* namely WRKY 15, 16, 19, 21, 27 the codon ending with TA is preferred. Kumar and Sharma.,1995 observed that G ending codons for threonine, alanine, proline, serine are avoided by *B.napus*, *B. oleracea*, and *B. campestris*. The WRKY genes of *Brassica rapa* also avoided G ending codons for threonine, proline, serine, and alanine, instead preference is given to codons ending with A, U, and C.

Table 1. Gene length, CAI value, percent GC content at first, second, third position, ENc value of *Arabidopsis thaliana* WRKY genes

Gene Name	Gene length	CAI value	%GC	%GC1	%GC2	%GC3	%GC12	ENC value
WRKY 1	1392	0.735	44.3	54.5	40.3	47.4	37.9	53.6
WRKY 2	2064	0.738	44.5	53.3	46.9	50.1	33.1	49
WRKY 3	1869	0.736	43.1	50.2	45.4	47.8	33.7	49.4
WRKY 4	1545	0.699	46.7	55.1	49.3	52.2	35.5	49.4
WRKY 6	3969	0.681	36	35.1	36.9	36	35.9	54.3
WRKY 7	1401	0.694	44.8	41.8	46.3	44.05	46.5	58.2
WRKY 8	981	0.714	42.2	46.5	38.5	42.5	41.6	50.8
WRKY 9	1125	0.734	42.3	48.5	40.5	44.5	37.9	51.1
WRKY 10	1458	0.69	44.5	52.1	41.2	46.65	40.3	50
WRKY 11	975	0.64	48	49.5	46.5	48	48	57.3
WRKY 12	657	0.67	41.9	42	44.7	43.35	38.8	48.9
WRKY 13	915	0.68	43.1	46.6	42.6	44.6	40	50.9
WRKY 14	1266	0.629	48.5	48.6	46.9	47.75	50	55.1
WRKY 15	1176	0.674	42.5	45.4	48	46.7	34.2	48.6
WRKY 16	4119	0.713	42.3	50.5	35.2	42.8	41.2	54.5
WRKY 18	1065	0.716	41.6	40.3	46.5	43.4	38	49.7
WRKY 19	5889	0.70	42.7	48.7	40.7	44.7	38.9	54.1
WRKY 20	1872	0.68	45.5	49.4	47.1	48.25	40.1	54.1
WRKY 21	1500	0.679	41.5	41.4	35.2	38.3	47.8	53
WRKY 22	897	0.641	48.4	49.8	52.2	51	43.1	56.4
WRKY 23	1014	0.683	44.4	50	41.7	44.15	41.4	53.7
WRKY 24	540	0.679	44.1	48.9	38.3	43.6	45	61
WRKY 25	1182	0.726	42	47.2	42.6	44.9	36	48.7
WRKY 26	1020	0.69	43	42.4	44.7	43.55	42.1	56.7
WRKY 27	1047	0.70	43.4	55	45.6	50.3	29.5	47.8
WRKY 28	957	0.704	42.9	47	45.5	46.25	36.4	50.5
WRKY 29	915	0.636	45.6	45.9	47.2	46.55	43.6	56
WRKY 30	912	0.726	43.4	48	40.8	49.85	41.4	49.3
WRKY 31	1617	0.684	49.3	54.2	45.5	49.85	48.2	56.3
WRKY 32	1704	0.658	44.1	37.1	51.9	44.5	43.1	54.3
WRKY 33	1539	0.728	43.1	42.7	48.1	45.4	38.6	45.7
WRKY 34	1704	0.698	46.6	51.7	47.1	49.4	40.9	56.6
WRKY 35	1284	0.653	48.6	46.7	49.3	48	49.8	56.6
WRKY 36	1164	0.682	43.4	48.2	41	44.6	41	56.9
WRKY 38	864	0.712	41.6	46.2	40.3	43.25	38.2	55.1

WRKY 39	993	0.734	42.8	44.7	46.2	45.45	37.5	52.4
WRKY 40	837	0.673	45	41.2	47	44.1	47	51.4
WRKY 41	942	0.636	44.9	43	46.5	44.75	45.2	55.3
WRKY 42	1587	0.685	47.3	52.7	44.2	48.45	45	57.1
WRKY 43	382	0.691	43.6	44.7	35.1	39.9	51.1	61
WRKY 44	1050	0.747	42.7	46.9	49.1	48	32	46.7
WRKY 45	444	0.699	47.1	47.3	40.5	43.9	53.4	46.6
WRKY 46	888	0.68	40.9	41.2	33.1	37.15	48.3	48.9
WRKY 47	1470	0.69	46.2	52.7	45.7	49.2	40.2	57.1
WRKY 48	1200	0.72	41.9	45.2	41.5	43.35	39	48.1
WRKY 49	894	0.745	36.5	40.6	33.2	36.9	35.6	57
WRKY 50	1080	0.719	39.3	37.5	43.1	40.3	37.2	52.3
WRKY 51	945	0.698	34.8	35.9	31.1	33.5	37.5	52.5
WRKY 53	1329	0.693	41.3	40.6	41.8	41.2	41.5	55.8
WRKY 54	1041	0.708	44.1	47.8	40.6	44.2	43.8	49.2
WRKY 55	1101	0.655	46	46.6	41.1	43.85	50.1	59.7
WRKY 56	588	0.734	41.2	48	39.8	43.9	35.7	51.6
WRKY 57	864	0.691	47.5	51.7	49.7	50.7	41	49.3
WRKY 58	1272	0.672	47	54.2	44.8	49.5	42	55
WRKY 59	609	0.691	40.4	39.9	39.4	39.65	41.9	56.4
WRKY 60	816	0.696	43.1	48.2	39.3	43.75	41.9	52.2
WRKY 61	1443	0.704	42.3	45.1	43	44.05	38.7	51.9
WRKY 62	909	0.659	47.2	40.9	58.4	49.65	42.2	55.9
WRKY 63	726	0.721	42	49.6	36.8	43.2	39.7	55.5
WRKY 64	750	0.724	41.6	49.2	36.8	43	38.8	51.1
WRKY 65	780	0.603	49.5	50.4	48.8	49.6	49.2	53.9
WRKY 66	819	0.68	41	36.3	39.6	37.95	47.3	56.2
WRKY 67	765	0.72	42.5	47.8	38	42.9	41.6	52.5
WRKY 69	966	0.65	46.2	49.7	44.4	47.05	44.4	57.32
WRKY 70	1083	0.63	41.2	37.7	42.9	40.30	42.9	52.90

Table 2. Gene length, CAI value, percent GC content at first,second,third position, ENc value of *Brassica rapa* WRKY genes

Gene Name	Gene length	CAI value	%GC	%GC1	%GC2	%GC3	%GC12	ENC value
WRKY 1	858	0.68	44.5	49.0	42.3	42.3	45.65	53.0
WRKY 2	2583	0.67	48.4	54.0	47.3	43.9	50.65	51.6
WRKY 3	1428	0.68	47.4	54.2	47.1	41.0	50.65	56.0
WRKY 4	1653	0.66	49.1	55.7	50.2	41.3	52.95	53.7
WRKY 6	1047	0.67	47.4	52.3	44.8	45.0	48.55	58.2
WRKY 7	1140	0.62	50.0	49.6	51.0	49.3	50.30	55.8
WRKY 8	987	0.70	43.0	48.2	41.3	39.5	44.75	52.7
WRKY 9	1965	0.70	43.3	50.2	40.4	39.2	44.50	56.1
WRKY 10	999	0.72	46.1	56.0	38.8	43.5	47.40	52.8
WRKY 11	657	0.57	51.7	49.2	48.9	56.8	49.05	55.8
WRKY 12	882	0.66	43.2	43.4	42.5	43.8	42.95	49.6
WRKY 13	1191	0.69	41.5	45.6	42.9	36.1	44.25	47.2
WRKY 14	960	0.65	48.0	50.1	46.6	47.4	48.35	57.6
WRKY 15	612	0.62	48.8	48.8	51.2	46.2	50.00	53.1
WRKY 16	612	0.67	44.8	49.5	40.7	44.1	45.1	44.1
WRKY 18	969	0.63	49.9	49.2	40.6	51.7	44.50	48.1
WRKY 19	2709	0.72	43.0	48.4	36.2	39.9	41.80	56.5
WRKY 20	1608	0.69	42.9	47.4	47.4	45.2	50.30	53.5
WRKY 21	1020	0.69	47.4	45.3	48.2	41.6	46.75	49.8
WRKY 22	897	0.57	46.4	48.5	53.5	45.6	51.0	51.1

WRKY 23	972	0.70	52.8	51.2	43.5	56.5	47.35	53.8
WRKY 24	531	0.70	44.2	48.0	37.3	38.0	42.65	46.8
WRKY 25	1122	0.68	42.0	51.6	42.0	40.7	46.80	51.3
WRKY 26	960	0.72	44.9	45.9	42.5	41.2	44.20	50.6
WRKY 27	1035	0.655	44.6	54.2	46.4	45.3	50.30	48.9
WRKY 28	939	0.68	46.5	46.3	43.8	38.8	45.05	52.7
WRKY 29	927	0.64	43.7	45.6	46.6	40.9	46.10	57
WRKY 30	942	0.72	45.0	50.0	43.8	42.7	44.45	51.8
WRKY 31	1560	0.67	43.8	51.7	46.6	42.7	48.45	58.3
WRKY 32	1380	0.68	48.1	55.4	38.9	47.3	50.75	54.7
WRKY 33	1557	0.63	47.8	45.1	45.2	42.0	47.40	44.6
WRKY 34	1650	0.68	43.7	51.3	46.1	36.2	48.40	58
WRKY 35	1212	0.63	46.4	46.5	49.7	42.5	48.15	59
WRKY 36	1152	0.70	48.8	48.4	45.5	50.2	44.65	55.6
WRKY 38	477	0.74	42.2	44.7	49.8	37.2	44.05	44.8
WRKY 39	1029	0.71	42.8	46.1	40.9	40.3	45.35	49.4
WRKY 41	996	0.65	43.1	45.5	43.4	38.8	45.05	53.2
WRKY 42	1560	0.67	46.2	51.7	44.6	48.5	48.45	58.3
WRKY 43	382	0.69	48.1	44.7	44.6	47.3	39.9	61.0
WRKY 44	1050	0.74	43.6	46.9	45.2	51.1	48.00	46.7
WRKY 45	435	0.69	42.7	46.2	35.1	32.0	41.70	49.2
WRKY 46	852	0.66	46.0	40.8	37.3	54.5	39.05	54.1
WRKY 47	1473	0.67	47.1	50.3	46.2	50.7	48.25	52.8
WRKY 48	1179	0.70	42.7	42.7	43.5	42.0	43.10	54.3
WRKY 49	987	0.71	40.4	43.2	35.9	42.2	39.55	54.9
WRKY 51	597	0.74	41.5	39.7	41.7	43.2	40.70	55.2
WRKY 53	972	0.62	48.3	50.0	45.7	49.1	47.85	56.9
WRKY 54	894	0.72	43.6	52.7	36.6	41.6	44.65	52.1
WRKY 55	855	0.66	49.0	50.9	45.6	50.5	48.25	61.0
WRKY 56	561	0.68	42.8	47.6	38.5	42.2	43.05	54.2
WRKY 57	885	0.68	47.0	51.2	50.8	39.0	51.00	51.5
WRKY 58	3249	0.67	46.9	51.9	46.3	42.6	49.10	56.9
WRKY 59	591	0.71	40.4	40.1	38.6	42.6	39.35	57.7
WRKY 60	927	0.69	42.6	48.2	39.8	39.8	44.00	49.6
WRKY 61	1197	0.69	43.5	46.1	43.4	41.1	44.75	53.6
WRKY 62	702	0.66	46.3	47.9	39.3	51.7	43.60	57.4
WRKY 64	777	0.68	46.2	52.5	37.8	48.3	45.15	55.6
WRKY 65	786	0.61	49.9	51.5	48.1	50.0	49.80	61.0
WRKY 66	1359	0.70	42.2	44.4	38.6	43.7	41.50	52.1
WRKY 67	831	0.70	44.0	49.1	38.3	44.8	44.85	54.8
WRKY 68	960	0.67	44.1	51.6	40.6	40.0	46.10	55.9
WRKY 69	849	0.63	51.4	54.8	46.6	52.7	50.70	57.2
WRKY 70	852	0.68	44.1	46.1	37.0	49.3	41.55	55.0

3.2 RSCU Analysis

The maximum RSCU value of 5 was found in *Brassica rapa* codon (UUU) coding for phenylalanine and the value 4.5 for codon (AGA) in *Arabidopsis thaliana*. Phenylalanine acts as building block of proteins and is also involved in secondary metabolite synthesis which helps in plant defense (Tzin & Galili., 2010). The codon coding for arginine was found out to be maximum in both of the species with RSCU value > 2. The codon AGA has been given higher preference as compared to AGG, CGG, CGA, CGC codons coding for arginine in most of the WRKY genes observed in *Arabidopsis thaliana* and *Brassica rapa*. Thirty *Brassica rapa* WRKY genes and twenty *Arabidopsis thaliana* WRKY genes possessed this higher RSCU values. The number of preferred codons in all the WRKY genes of both the species were different. A maximum number of preferred codons in *Arabidopsis thaliana* is twenty-six and the minimum is twelve. WRKY 57 has the maximum number of preferred

codon *i.e.* twenty-six. In the case of *Brassica rapa*, the maximum number of preferred codons was twenty-three and the minimum was thirteen. WRKY 22 and WRKY 51 has a maximum of twenty-three preferred codons.

3.3 ENc Analysis

The ENc values of two plant species WRKY genes were observed, the range of value were from 44 to 61 in *Brassica rapa* and 45.7 to 61 in *Arabidopsis thaliana*. The observed ENc values are all greater than 40, this reveals that codon usage in WRKY genes show a lesser amount of biases. Highest biasness is shown by WRKY 33 in both the plant species with ENc value of 45.7 in case of *Arabidopsis thaliana* and 44.6 in case of *Brassica rapa*. While some of the genes WRKY 45, 55, 65 showed no CUB in *Brassica rapa*. Similarly, WRKY 43 and 55 showed no biases in *Arabidopsis thaliana*.

3.4 CAI Analysis

After comparing the CAI values of WRKY genes of both *Arabidopsis thaliana* and *Brassica rapa* it was found out that the CAI values of *Arabidopsis thaliana* were high, however, some of the WRKY genes of *Brassica rapa* namely WRKY 10, 13, 14, 20, 21, 23, 24, 26, 28, 30, 31, 33, 34, 39, 41, 43, 64, 67, 70 showed higher CAI values as compared to *Arabidopsis thaliana*, depicting there higher expression level. The range of the CAI value was from 0.5 to 0.7 (Tables 1 and 2). The higher is the CAI value higher is the expression level of genes. WRKY genes are an important group of plant transcription factors and it has been proposed that highly expressed genes, including ribosomal proteins and transcription and translation factors, tend to have high CAI (Post & Nomura, 1980; Gouy & Gautier, 1982).

3.5 ENc and GC$_3$ Plot

The Nc plot (ENc versus GC$_{3s}$) was examined to know about the pattern of codon usage (Wright, 1990). The Figures 1A and 1B show the distribution plot of ENc against GC3s for WRKY genes of *Arabidopsis thaliana* and *Brassica rapa*. The solid blue dots depicts the expected position of genes whose codon usage pattern is only determined by the GC3s. In Figures 1A and 1B, we can see that majority of the points are lying below the expected curve this indicates that some other factors independent of nucleotide composition affect the codon usage variation among the WRKY genes of *Arabidopsis thaliana* and *Brassica rapa*. Most of the WRKY gene CDS in the two plant species appeared to cluster around the expected ENc of 30% to 50%. The ENc and GC$_3$ plot of the two plant species depicts that natural selection has influenced the CUB in WRKY transcription factor genes.
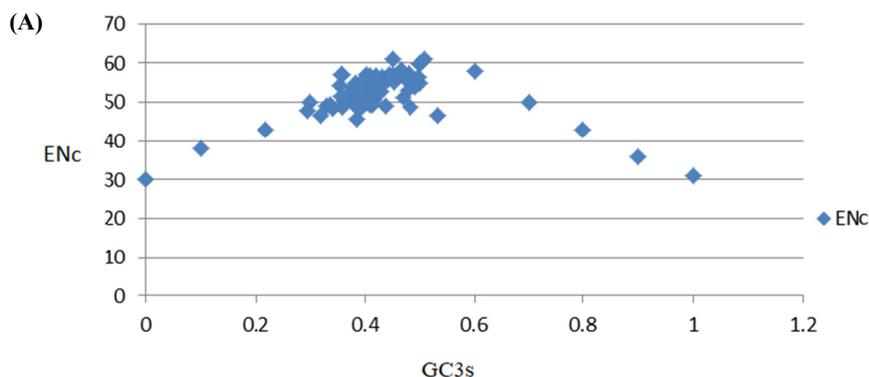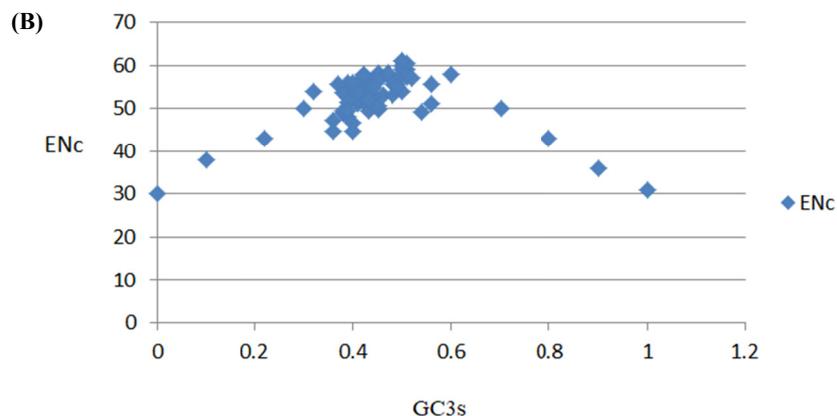
Figure 1. (A) ENc versus GC3s plot of WRKY genes of *Arabidopsis thaliana* (B) ENc versus GC3s plot of WRKY genes of *Brassica rapa*. The bell shaped curve shows the expected ENc value if the codon bias is only due to GC3s

3.6 Hierarchical Clustering of WRKY Genes

All the WRKY genes of *Arabidopsis thaliana* and *Brassica rapa* were analyzed using the clustered heat map analysis tool by giving RSCU values using the hierarchical clustering algorithm. Among the WRKY genes within the same species, a close relationship was observed between the genes which shared similar codon usage. Species with a close genetic relationship always show a similar codon usage pattern (Sharp et al., 1988). Liu et al., 2005 worked on *Oryza sativa* and found that genes having involvement in metabolic processes have a preference for C or G in GC_3s position stating about the strong correlation between synonymous codon usage and gene function. Based on the above studies it can be said that in case of *Arabidopsis thaliana* (Figure 2A) and *Brassica rapa* (Figure 2B) the WRKY genes that come under the same cluster have similar RSCU code and might have similar functions. The similarity in gene functions of some of the WRKY genes has been well documented by Kayum et al. (2005). WRKY gene 1 and 18 of *Brassica rapa* come in the same cluster and both the genes are responsive towards fungal pathogen, abscisic acid, 6-Benzylaminopurine, jasmonic acid, salicylic acid and ethylene (Yang et al., 2009). Similarly, WRKY gene 2 and 47 of *Brassica rapa* also comes under the same cluster, have similar RSCU code and shown to respond against drought, salinity, phosphorus, and nitrogen starvation (Xiong et al., 2013).
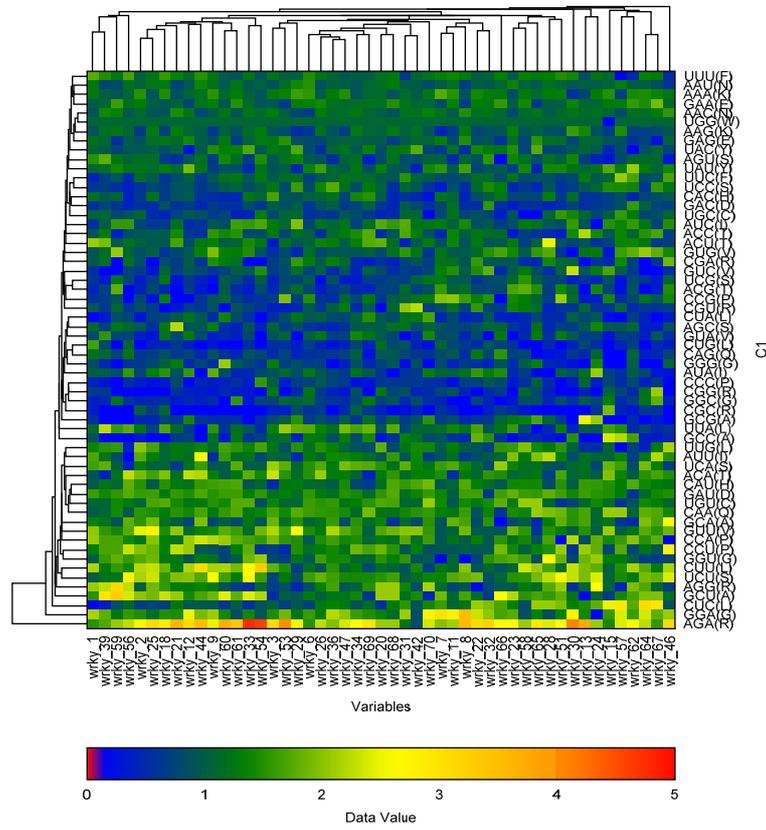
Figure 2A. Cluster tree based on the RSCU value of *Arabidopsis thaliana* WRKY genes
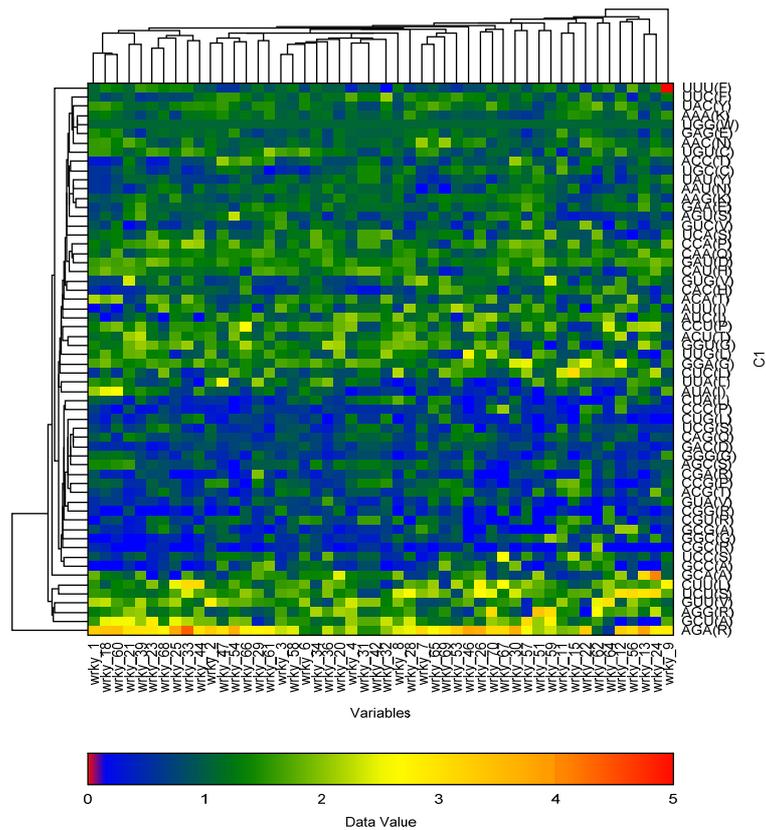

Figure 2B. Cluster tree based on the RSCU value of *Brassica rapa* WRKY genes

Clustering of relative synonymous codon usage (RSCU) values of each codon among WRKY genes of (A) *Arabidopsis thaliana* (B) *Brassica rapa*. Each rectangular box on the map represents the RSCU values of codon (shown in rows) across WRKY genes. The intensity of color coding indicates different RSCU values.

3.7 Neutrality Plot

Neutrality plot was drawn with GC_{12} as Y-axis and GC_3 as X-axis. The dot represents each WRKY gene of the particular species. This plot helps us to know about the key determining factors that shaped codon usage (Kumar et al., 2016). When the slope of the regression line is equal to 1, a perfect correlation is said to occur between GC_{12} and GC_3 and mutation pressure is a dominant factor resulting in CUB. In (Figure 3A) the observation of neutrality plots of WRKY genes CDS in *Arabidopsis thaliana* shows the negative (-0.08) slope of regression line while the slope was zero in *Brassica rapa* (0.03) this indicates that the influence of direct mutation pressure for CUB is only 3% in *Arabidopsis thaliana* while the impact of natural selection was calculated out to be 97%. The genes showed low mutation bias level and it can be said that natural selection is a dominant force acting in shaping the codon usage pattern of WRKY genes of *Brassica rapa* and *Arabidopsis thaliana*. Highly expressed genes, such as translation elongation factors and ribosomal proteins have natural selection acting on them to ensure efficient translation (Hershberg & Petrov, 2008).

(A)   (B)
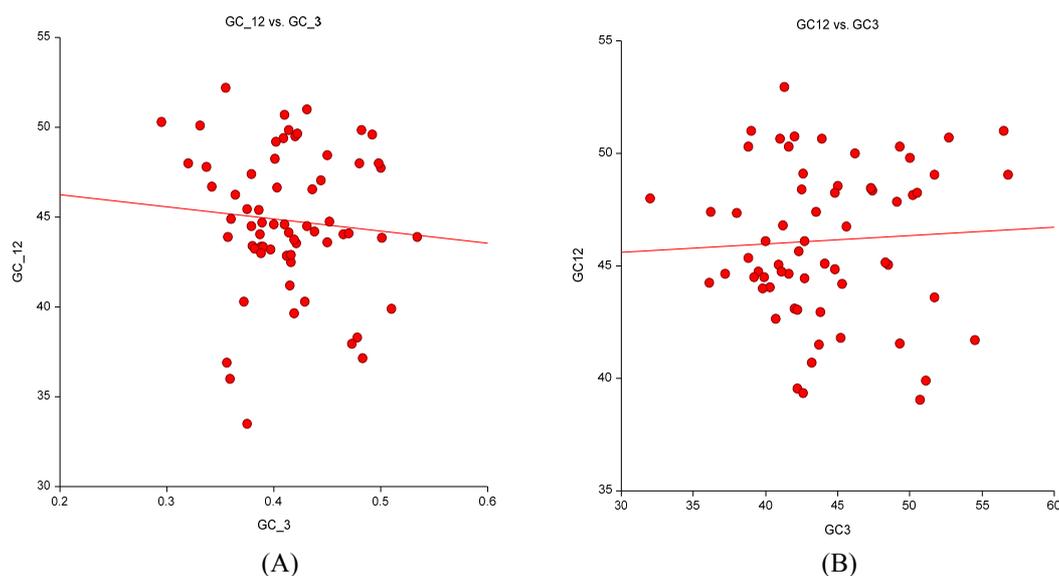
Figure 3. Neutrality plot (GC12 against GC3) (A) *Arabidopsis thaliana*, the regression line is y = 0.063x + 0.502, R^2 = 0.0072; (B) *Brassica rapa*, y = 0.143x + 0.401, R^2 = 0.0031

3.8 Effect of Gene Expression Level and Protein Length on Codon Usage Bias

In the (Figures 4A and 4B) and (Figures 5A and 5B) the significant negative correlation between CAI and GC3 (R^2 = 0.3, $p < 0.05$, slope = -0.4) and (R^2 = 0.19, $p < 0.05$, slope = -0.5) were shown in case of *Arabidopsis thaliana* and *Brassica rapa* WRKY genes respectively. In Figures 6A and 6B, positive correlation was observed between protein length and ENc (R^2 = 0.004, $p > 0.05$, slope = 0.003) in *Arabidopsis thaliana* and *Brassica rapa* WRKY genes respectively. We can conclude that protein length affects codon usage bias in WRKY genes of the two plant species.

(A)   (B)
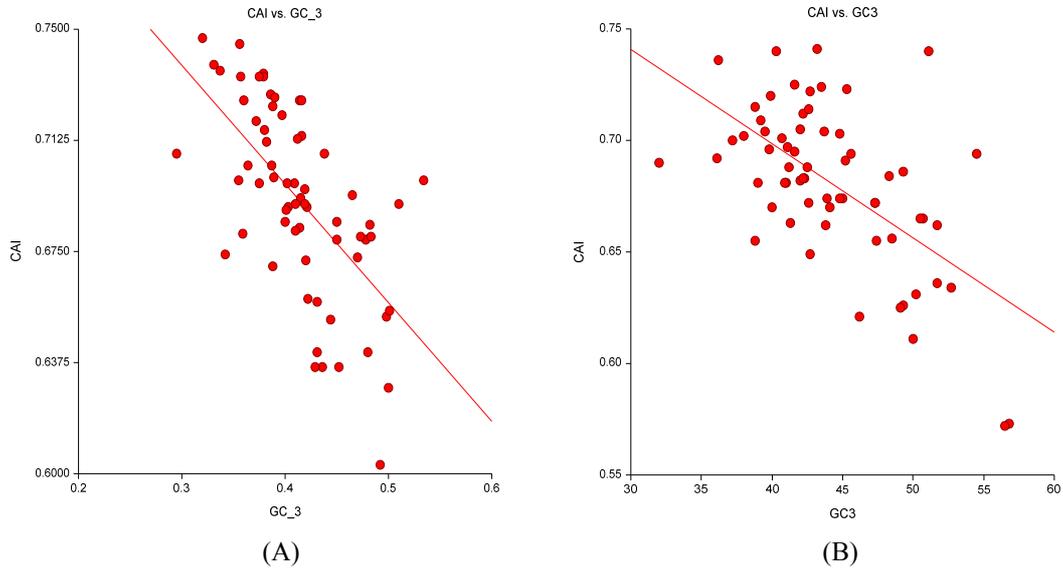
Figure 4. CAI versus GC3 plot of (A) *Arabidopsis thaliana* WRKY genes, $y = 0.4x + 0.85$, $R^2 = 0.39$ and (B) *Brassica rapa* WRKY genes, $y = 0.042x + 0.86$, $R^2 = 0.35$


(A)   (B)
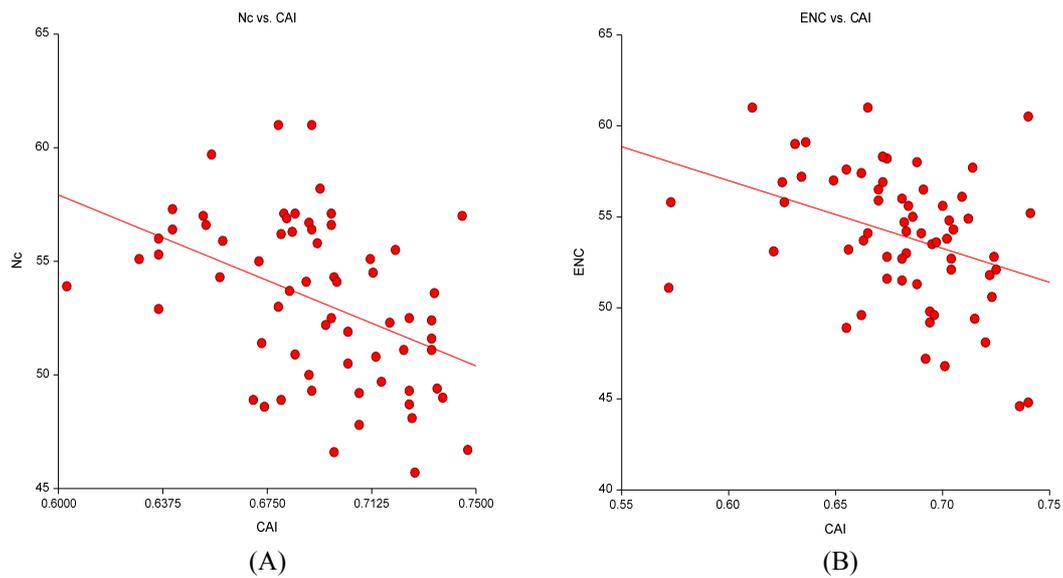
Figure 5. ENc versus CAI plot of (A) *Arabidopsis thaliana* WRKY genes, $y = -0.657x + 48.007$, $R^2 = -0.19$ (B) *Brassica rapa* WRKY genes, $y = -0.543x + 46.005$, $R^2 = -0.23$

(A)   (B)
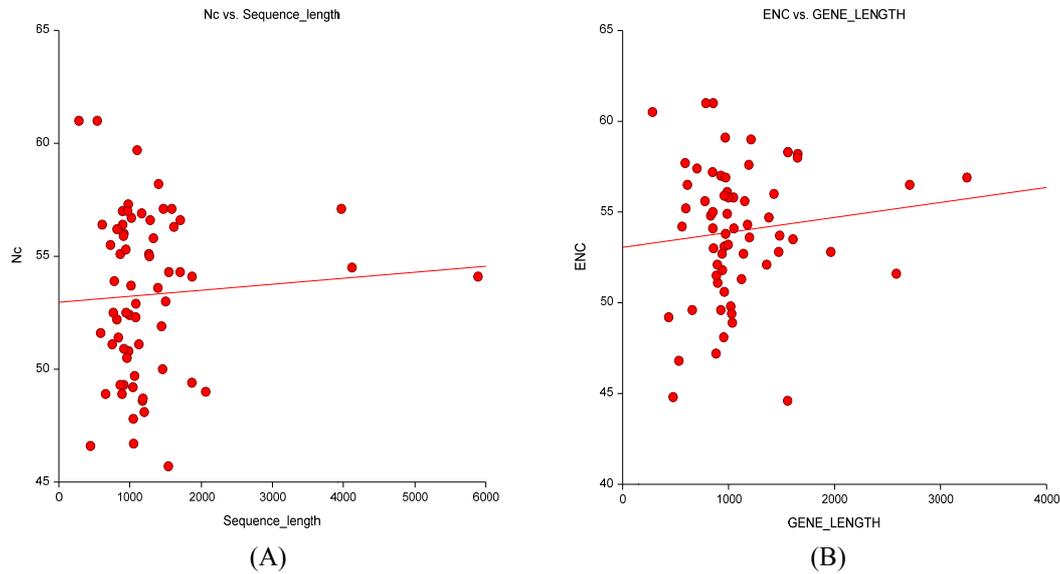
Figure 6. ENc versus Gene length plot of (A) *Arabidopsis thaliana* WRKY genes, $y = 2.54x + 23$, $R^2 = 0.0041$, (B) *Brassica rapa* WRKY genes, $y = 4.54x + 14$, $R^2 = 0.01$

3.9 Correspondence Analysis

For investigating the variation of synonymous codon usage pattern in *Arabidopsis thaliana* and *Brassica rapa* WRKY genes. Correspondence analysis was conducted using RSCU values (He et al., 2016). The axes factors are shown in Figures 7A and 7B where axes F1 and F2 correlate to the two main influencing factors of CUB. In the case of *Arabidopsis thaliana,* they represent 19.86% and 15.10% of variation respectively (Figure 7A). Similarly in the case of *Brassica rapa* (Figure 7B) the axes factor F1 and F2 represent 19.60% and 16.49% of variation respectively. This suggests that the first axis is the primary factor of the overall variation. All the genes were diffusely distributed this indicates that many factors affect CUB (Zhao et al., 2016). In *Arabidopsis thaliana* correspondence analysis plot of WRKY genes, the distribution density of triplet codon ending with A and G is more towards axes F1 than the codon ending with A and C (Figure 7A). Similarly as shown in (Figure 7B) the *Brassica rapa* correspondence analysis plot of WRKY genes the A and G ending codons are more towards axes F1 . These results propose that nucleotide composition has an influence on the CUB of WRKY genes in *Arabidopsis thaliana* and *Brassica rapa*.


(A)   (B)
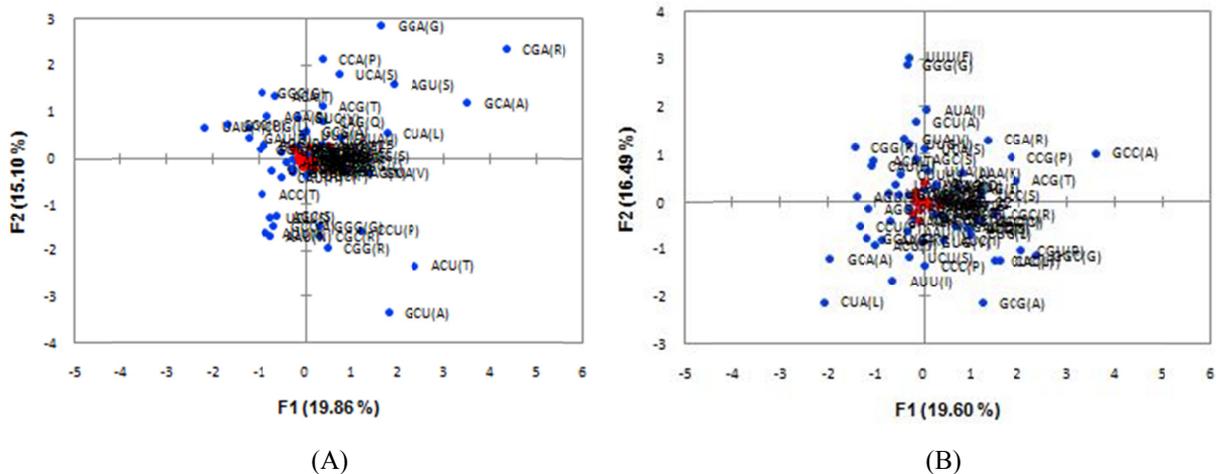
Figure 7. Correspondence analysis of relative synonumous codon usage (RSCU) for WRKY genes of (A) *Arabidopsis thaliana* (B) *Brassica rapa*. F1 and F2 are described as the two factors that are responsible for the variation in the synonymous codons. The blue, yellow and black dots indicate the codons used by the amino acids

4. Discussion

Our analysis of codon usage patterns among the WRKY genes of *Arabidopsis thaliana* and *Brassica rapa* reveals that most of the WRKY genes in both the species have higher AT content as compared to GC content the reason might be the small size of a protein sequence as shorter sequences have higher AT bias (Wuitschick & Karrer, 1999). The G+C content was lowest at GC_3 codon position and the status of G and C ending codons were found to be high in some of the WRKY genes in *Arabidopsis thaliana* and *Brassica rapa.* These results suggest that selection has driven codon usage of genes having an important function to have high GC content. Many factors affect the codon usage pattern among and within the species. In highly expressed genes codons that improve translational efficiency are selected while in the genes that have a low level of expression mutation-drift determines the codon usage (Bulmer, 1991). In the study, the ENC versus GC3 plot of WRKY genes of *Arabidopsis thaliana* and *Brassica rapa* showed that natural selection influences the major codon usage pattern. Similar results about WRKY genes of *Medicago truncatula* were obtained by Song et al. (2015). The gene expression level of different genes must be known if the analysis is done to know about the relationship between codon bias and gene expression level. In eukaryotes, it is difficult to find out the gene expression level as there is the difference in expression at different times and different tissues. In our study, we use CAI to evaluate the expression level of WRKY genes. CAI has now been considered as a well-accepted measure of gene expression (Naya et al., 2001; Gupta et al., 2004). The hierarchical clustering grouped the WRKY genes with similar RSCU values within the same cluster, the genes present in the same cluster showed similar functions. The presence of codons AGA coding for arginine with RSCU value greater than 2 was seen in all the WRKY gene sequences analyzed. The preferential use of arginine can be linked to oxidative stress response. A study performed in budding yeast revealed about the reprogramming of tRNA, which leads to the codon biased mRNA coding for arginine to be expressed under oxidative stress conditions (Gu et al., 2014). Arginine codons also plays a role in the evolution and the variability of Hepatitis A virus strains (Andrea et al., 2011). Insights into the study of synonymous codons usage pattern of WRKY genes in *Arabidopsis thaliana* and *Brassica rapa* is provided by our study. After comparison of RSCU values of different amino acids of *Arabidopsis thaliana* and *Brassica rapa* WRKY genes, it was seen that some WRKY genes have a small difference in the codon usage and hence can be used in transgenic studies.

5. Conclusion

In order to know regarding the forces that are responsible for CUB of WRKY transcription factor genes in two related species *Arabidopsis thaliana* and *Brassica rapa*, we examined the WRKY coding sequences of both the plant species that were present in the database. Different indices that help in the prediction of CUB have been calculated and on the basis of CAI values, it was investigated that WRKY genes are highly expressed genes that has higher AT content as compared to GC content and they show a moderate level of biases in both the plant species. Natural selection acts as main determining force shaping the codon usage pattern. WRKY genes having similar RSCU values were found to share similar functions. A positive correlation is seen between the coding sequence length of WRKY genes and effective number of codons.

Acknowledgements

References

Akashi, H. (1997). Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. *Gene, 205*(1), 269-278. https://doi.org/10.1016/S0378-1119(97)00400-9

Behura, S. K., & Severson, D. W. (2013). Codon usage bias: Causative factors, quantification methods, and genome-wide patterns: With emphasis on insect genomes. *Biological Reviews, 88*(1), 49-61. https://doi.org/10.1111/j.1469-185X.2012.00242.x

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics, 129*, 897-907.

Campbell, W. H., & Gowri, G. (1990). Codon usage in higher plants, green algae, and cyanobacteria. *Plant Physiology, 92*(1), 1-11. https://doi.org/10.1104/pp.92.1.1

Coulondre, C., Miller, J. H., Farabaugh, P., & Gilbert, J. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature, 274*(5673), 775-780. https://doi.org/10.1038/274775a0

Chen, H., Sun, S., Norenburg, J. L., & Sundberg, P. (2014). Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms (Nemertea). *PloS One, 9*(1), e85631. https://doi.org/10.1371/journal.pone.0085631

Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences, 96*(8), 4482-4487. https://doi.org/10.1073/pnas.96.8.4482

D'Andrea, L., Pintó, R. M., Bosch, A., Musto, H., & Cristina, J. (2011). A detailed comparative analysis on the overall codon usage patterns in hepatitis A virus. *Virus Research, 157*(1), 19-24. https://doi.org/10.1016/j.virusres.2011.01.012

Eulgem, T., Rushton, P. J., Robatzek, S., & Somssich, I. E. (2000). The WRKY superfamily of plant transcription factors. *Trends in Plant Science, 5*(5), 199-206. https://doi.org/10.1016/S1360-1385(00)01600-9

Goodarzi, H., Torabi, N., Najafabadi, H. S., & Archetti, M. (2008). Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons. *Gene, 407*(1), 30-41. https://doi.org/10.1016/j.gene.2007.09.020

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research, 10*(22), 7055-7074. https://doi.org/10.1093/nar/10.22.7055

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*.

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology, 22*(7), 346-353. https://doi.org/10.1016/j.tibtech.2004.04.006

Gu, C., Begley, T. J., & Dedon, P. C. (2014). tRNA modifications regulate translation during cellular stress. *FEBS Letters, 588*(23), 4287-4296. https://doi.org/10.1016/j.febslet.2014.09.038

Gupta, S. K., Bhattacharyya, T. K., & Ghosh, T. C. (2004). Synonymous codon usage in Lactococcuslactis: Mutational bias versus translational selection. *Journal of Biomolecular Structure and Dynamics, 21*(4), 527-535. https://doi.org/10.1080/07391102.2004.10506946

Hambuch, T. M., & Parsch, J. (2005). Patterns of synonymous codon usage in Drosophila melanogaster genes with sex-biased expression. *Genetics, 170*(4), 1691-1700. https://doi.org/10.1534/genetics.104.038109

Hu, J., Zhao, X., Zhang, Z., & Yu, J. (2007). Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Research in Microbiology, 158*(4), 363-370. https://doi.org/10.1016/j.resmic.2007.02.007

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annu. Rev. Genet., 42*, 287-299. https://doi.org/10.1146/annurev.genet.42.110807.091442

Kayum, M. A., Jung, H. J., Park, J. I., Ahmed, N. U., Saha, G., Yang, T. J., & Nou, I. S. (2015). Identification and expression analysis of WRKY family genes under biotic and abiotic stresses in *Brassica rapa*. *Molecular Genetics and Genomics, 290*(1), 79-95. https://doi.org/10.1007/s00438-014-0898-1

Kawabe, A., & Miyashita, N. T. (2003). Patterns of codon usage bias in three dicots and four monocot plant species. *Genes & Genetic Systems, 78*(5), 343-352. https://doi.org/10.1266/ggs.78.343

Kumar, N., Bera, B. C., Greenbaum, B. D., Bhatia, S., Sood, R., Selvaraj, P., & Virmani, N. (2016). The revelation of influencing factors in overall codon usage bias of equine influenza viruses. *PloS One, 11*(4), e0154376. https://doi.org/10.1371/journal.pone.0154376

Kumar, P. A., & Sharma, R. P. (1995). Codon usage in Brassica genes. *Journal of Plant Biochemistry and Biotechnology, 4*(2), 113-115. https://doi.org/10.1007/BF03262965

Liu, Q., Dou, S., Ji, Z., & Xue, Q. (2005). Synonymous codon usage and gene function are strongly related in *Oryza sativa*. *Biosystems, 80*(2), 123-131. https://doi.org/10.1016/j.biosystems.2004.10.008

Mandlik, V., Shinde, S., & Singh, S. (2014). Molecular evolution of the enzymes involved in the sphingolipid metabolism of Leishmania: Selection pressure in relation to functional divergence and conservation. *BMC Evolutionary Biology, 14*(1), 142. https://doi.org/10.1186/1471-2148-14-142

Naya, H., Romero, H., Carels, N., Zavala, A., & Musto, H. (2001). Translational selection shapes codon usage in the GC-rich genome of Chlamydomonas reinhardtii. *FEBS Letters, 501*(2-3), 127-130. https://doi.org/10.1016/S0014-5793(01)02644-8

Ossowski, S. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana. Science, 327*(5961), 92-94. https://doi.org/10.1126/science.1180677

Post, L. E., & Nomura, M. (1980). DNA sequences from the str operon of *Escherichia coli*. *Journal of Biological Chemistry, 255*(10), 4660-4666.

Phukan, U. J., Jeena, G. S., & Shukla, R. K. (2016). WRKY transcription factors: molecular regulation and stress responses in plants. *Frontiers in Plant Science, 7*, 760. https://doi.org/10.3389/fpls.2016.00760

Puigbò, P., Bravo, I. G., & Garcia-Vallve, S. (2008). CAIcal: A combined set of tools to assess codon usage adaptation. *Biology Direct, 3*(1), 38. https://doi.org/10.1186/1745-6150-3-38

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution, 24*(12), 28-38. https://doi.org/10.1007/BF02099948

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., & Wright, F. (1988). Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomycespombe, Drosophila melanogaster and Homo sapiens: A review of the considerable within-species diversity. *Nucleic Acids Res., 16*, 8207-8211. https://doi.org/10.1093/nar/16.17.8207

Smith, H. B. (2000). *Signal transduction in systemic acquired resistance.*

Song, H., Wang, P., Ma, D., Xia, H., Zhao, C., Zhang, Y., & Zhao, S. (2015). Analysis of codon usage bias of WRKY transcription factors in Medicagotruncatula. *Journal of Agricultural Biotechnology, 23*(2), 203-212.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences, 85*(8), 2653-2657. https://doi.org/10.1073/pnas.85.8.2653

Supek, F., & Vlahoviček, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics, 6*(1), 182. https://doi.org/10.1186/1471-2105-6-182

Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B., & Feldmann, K. A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics, 11*, 308.

Tzin, V., & Galili, G. (2010). The biosynthetic pathways for shikimate and aromatic amino acids in *Arabidopsis thaliana*. *The Arabidopsis book/American Society of Plant Biologists, 8*.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene, 87*(1), 23-29. https://doi.org/10.1016/0378-1119(90)90491-9

Wang, H., Liu, S., Zhang, B., & Wei, W. (2016). Analysis of synonymous codon usage bias of Zika virus and its adaption to the hosts. *PloS One, 11*(11), e0166260. https://doi.org/10.1371/journal.pone.0166260

Wu, Y. Q., Li, Z. Y., Zhao, D. Q., & Tao, J. (2017). *Comparative analysis of flower-meristem-identity gene APETALA2 (AP2) codon in different plant species*.

Wuitschick, J. D., & Karrer, K. M. (1999). Analysis of genomic G+C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila. *Journal of Eukaryotic Microbiology, 46*(3), 239-247. https://doi.org/10.1111/j.1550-7408.1999.tb05120.x

Xiong, W., Xu, X., Zhang, L., Wu, P., Chen, Y., Li, M., Jiang, H., & Wu, G. (2013). Genome-wide analysis of the WRKY gene family in physic nut (*Jatropha curcas* L.). *Gene, 524*, 124-132. https://doi.org/10.1016/j.gene.2013.04.047

Yang, H. Q., & Gao, H. J. (2007). Physiological function of arginine and its metabolites in plants. *Journal of Plant Physiology and Molecular Biology, 33*(1), 1-8.

Yang, B., Jiang, Y., Rahman, M. H., Deyholos, M. K., & Kav, N. N. (2009). Identification and expression analysis of WRKY transcription factor genes in canola (*Brassica napus* L.) in response to fungal pathogens and hormone treatments. *BMC Plant Biology, 9*(1), 68. https://doi.org/10.1186/1471-2229-9-68

Yang, X., Ma, X., Luo, X., Ling, H., Zhang, X., & Cai, X. (2015). Codon usage bias and determining forces in Taenia solium genome. *The Korean Journal of Parasitology, 53*(6), 689. https://doi.org/10.3347/kjp.2015.53.6.689

Zhao, Y., Zheng, H., Xu, A., Yan, D., Jiang, Z., Qi, Q., & Sun, J. (2016). Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution. *BMC Genomics, 17*(1), 677. https://doi.org/10.1186/s12864-016-3021-7

Copyrights